Karlsruhe Institute of Technology

Probability & Computing

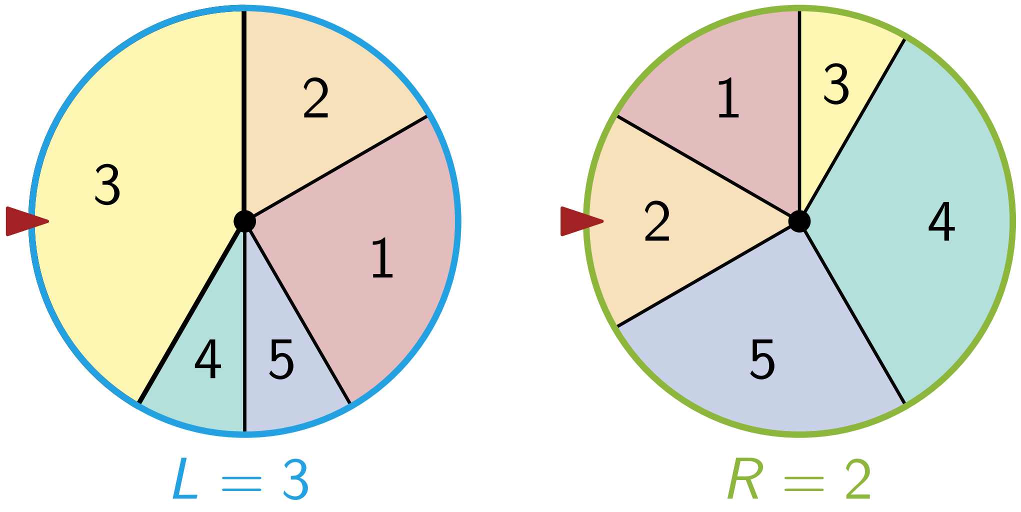
Coupling & Erdős-Rényi Random Graphs

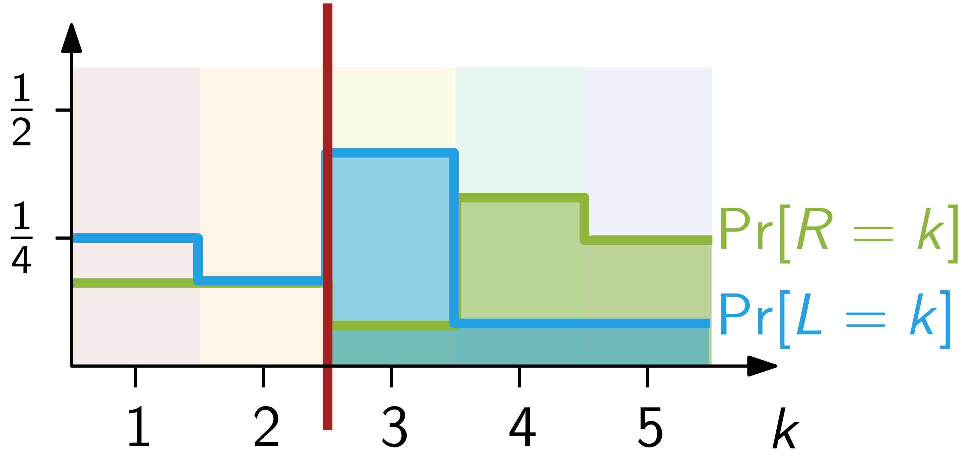
Wheels of Fortune

The Problem
- Consider the two wheels of fortune
- The higher the value the larger the price
- Which do you spin? Why? Can we prove that?

The Maths
- Let L be the value of the left wheel
- Let R be the value of the right wheel
- To show: For all values k: $\Pr[R \geq k] \geq \Pr[L \geq k]$

Proof
- For each k
  - Compute the sums of the probabilities
  - Compare
  - Tedious...





Wheels of Fortune

The Problem

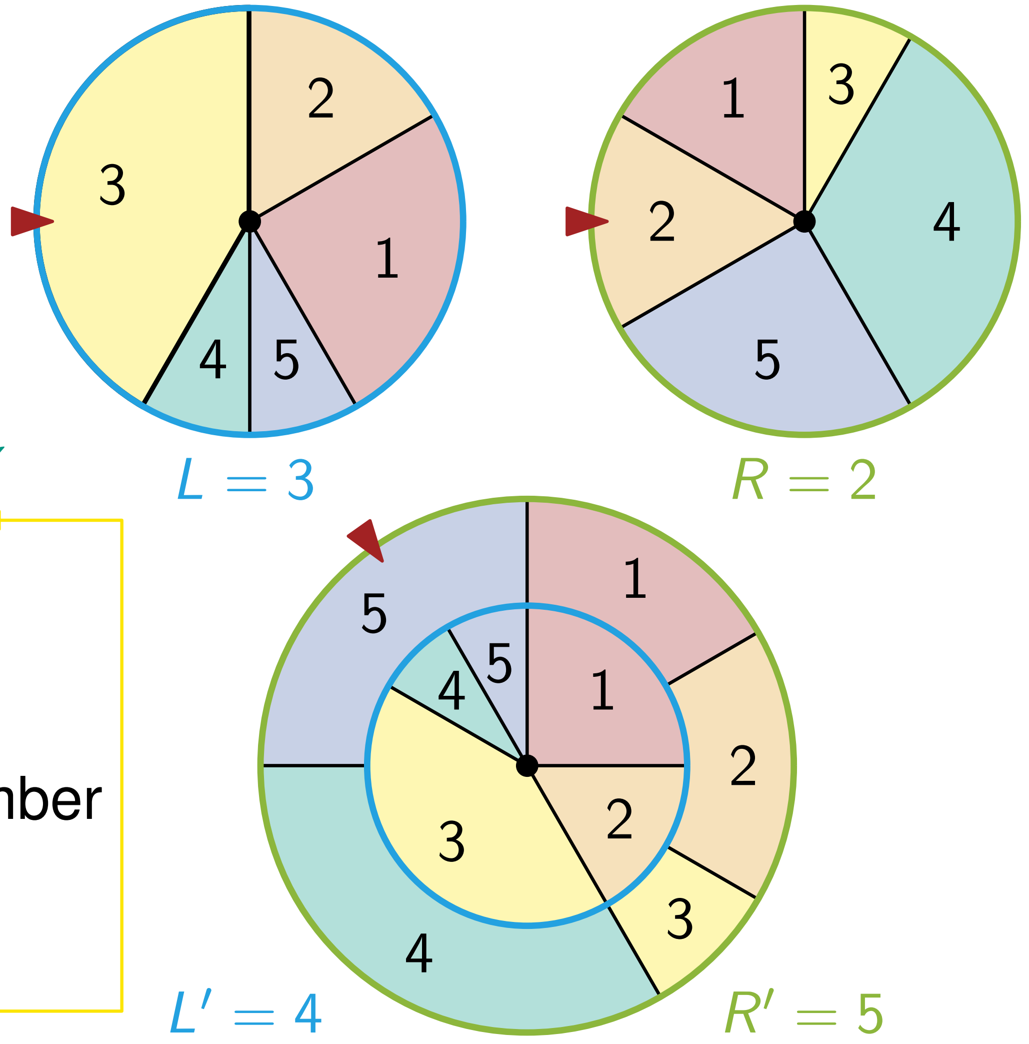
- Consider the two wheels of fortune
- The higher the value the larger the price
- Which do you spin? Why? Can we prove that?

The Maths

- Let L be the value of the left wheel
- Let R be the value of the right wheel
- To show: For all values k: $\Pr[R \geq k] \geq \Pr[L \geq k]$ ✓

Proof: Frankenstein's Wheel of Fortune!

- Sort the wheels (does not change their distributions)
- Adjust sizes and glue together
- Spin as one wheel: L' inner number, R' outer number
- Note that $L \overset{d}{=} L'$ and $R \overset{d}{=} R'$ equal distributions
- But L' and R' are dependent and *always* $R' \geq L'$
  $\Rightarrow \Pr[R' \geq k] \geq \Pr[L' \geq k]$



What just happened?

Setup & Method

independent

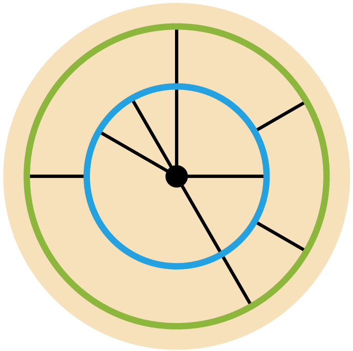
- Random variable L on the left wheel and R on right wheel

$L \overset{d}{=} L'$ $\quad$ $R \overset{d}{=} R'$

- Random variable L' on inner wheel and R' on outer wheel

dependent

The coupling defines how L' and R' are related



- Define a relation between random variables to make statements about one using the other Here: $\Pr[R' \geq k] \geq \Pr[L' \geq k] \Rightarrow \Pr[R \geq k] \geq \Pr[L \geq k]$

Definition: Let X_1, X_2 be random variables defined on probability spaces $(\Omega_1, \Sigma_1, \Pr_1)$ and $(\Omega_2, \Sigma_2, \Pr_2)$, respectively. A **coupling** of X_1 and X_2 is a pair of random variables (X_1', X_2') defined on a new probability space $(\Omega, \Sigma, \Pr)$ such that $X_1 \overset{d}{=} X_1'$ and $X_2 \overset{d}{=} X_2'$.

- X_1' and X_2' live in the *same* space
- Typically we define X_1' and X_2' to be dependent
- Typically we do not talk about the probability spaces explicitly

- Abstracting away technicalities, people just "couple" X_1 and X_2 "directly", without introducing X_1' and X_2'

Application: Biased Coins

The Problem

- We have a fair $\{0,1\}$-coin that yields 1 with probability $\frac{1}{2}$
- And an unfair $\{0,1\}$-coin that yields 1 with probability $\frac{2}{3}$
- Throw each coin n times, count the 1s, yielding F and U
- You pick a coin. You win if your coin gets more 1s than the other. Which do you pick?

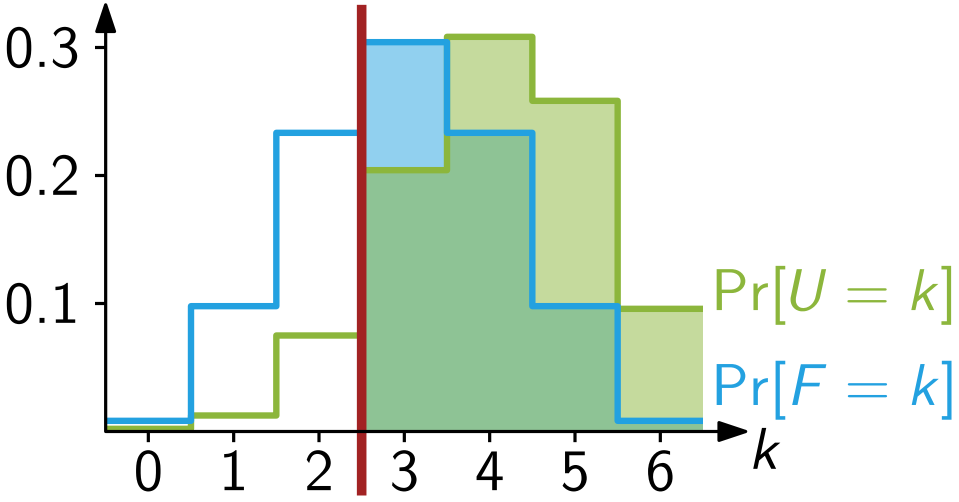
$F = \sum$ ⓪ ① ⓪ ⓪ ⓪ ① $= 2$

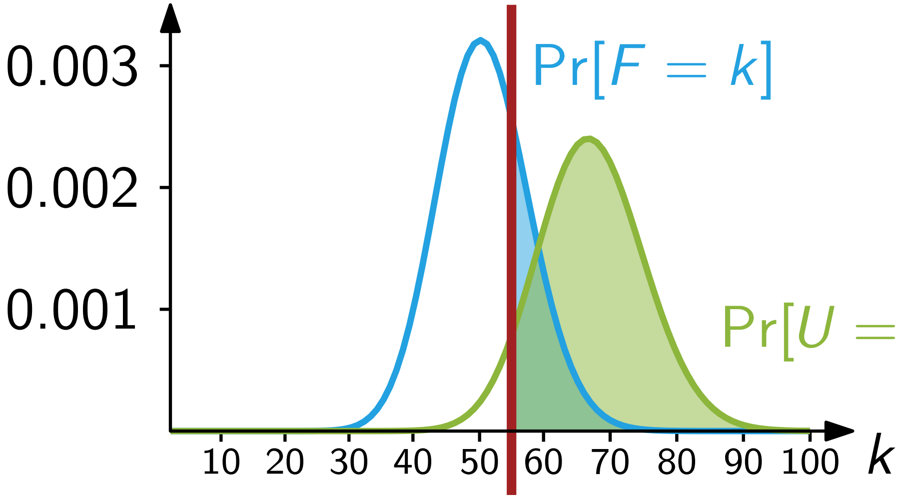
$U = \sum$ ⓪ ⓪ ① ① ① ① $= 4$

Claim $\Pr[U \geq k] \geq \Pr[F \geq k]$

Proof *Compare sums for all* $k \leq 6$

And if $n = 100$? *so many sums...*





Application: Biased Coins

The Problem

- We have a fair $\{0,1\}$-coin that yields 1 with probability $\frac{1}{2}$
- And an unfair $\{0,1\}$-coin that yields 1 with probability $\frac{2}{3}$
- Throw each coin n times, count the 1s, yielding F and U
- You pick a coin. You win if your coin gets more 1s than the other. Which do you pick?

$$F = \sum \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 = 2$$

$$U = \sum \ 0 \ 0 \ 1 \ 1 \ 1 \ 1 = 4$$

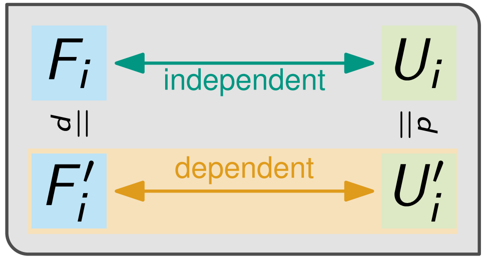
Claim $\Pr[U \geq k] \geq \Pr[F \geq k]$

Proof

- Let F_i be indicator for ith fair coin
- Let U_i be indicator for ith unfair coin
- Let W_i be the result of a fair die-roll
  - Define $F'_i = 1$ iff $W_i \leq 3 \Rightarrow F_i \stackrel{d}{=} F'_i$
  - Define $U'_i = 1$ iff $W_i \leq 4 \Rightarrow U_i \stackrel{d}{=} U'_i$
- F'_i and U'_i are dependent and *always* $U'_i \geq F'_i$

Coupling: Random variables X_1, X_2. Define random variables X'_1, X'_2 in a shared probability space such that $X_1 \stackrel{d}{=} X'_1$ and $X_2 \stackrel{d}{=} X'_2$.

W_i	1	2	3	4	5	6
Pr	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
F'_i	1	1	1	0	0	0
	$\wedge\vert$	$\wedge\vert$	$\wedge\vert$	$\wedge\vert$	$\wedge\vert$	$\wedge\vert$
U'_i	1	1	1	1	0	0



Application: Biased Coins

The Problem

- We have a fair $\{0,1\}$-coin that yields 1 with probability $\frac{1}{2}$
- And an unfair $\{0,1\}$-coin that yields 1 with probability $\frac{2}{3}$
- Throw each coin n times, count the 1s, yielding F and U

$F = \sum 0\,1\,0\,0\,0\,1 = 2$

$U = \sum 0\,0\,1\,1\,1\,1 = 4$

- You pick a coin. You win if your coin gets more 1s than the other. Which do you pick?

Claim $\Pr[U \geq k] \geq \Pr[F \geq k]$ ✓

Proof

- Let F_i be indicator for ith fair coin
- Let U_i be indicator for ith unfair coin
- Let W_i be the result of a fair die-roll
  - Define $F'_i = 1$ iff $W_i \leq 3 \Rightarrow F_i \overset{d}{=} F'_i$
  - Define $U'_i = 1$ iff $W_i \leq 4 \Rightarrow U_i \overset{d}{=} U'_i$
- F'_i and U'_i are dependent and *always* $U'_i \geq F'_i$
  $\Rightarrow U' \geq F' \Rightarrow \Pr[U' \geq k] \geq \Pr[F' \geq k]$

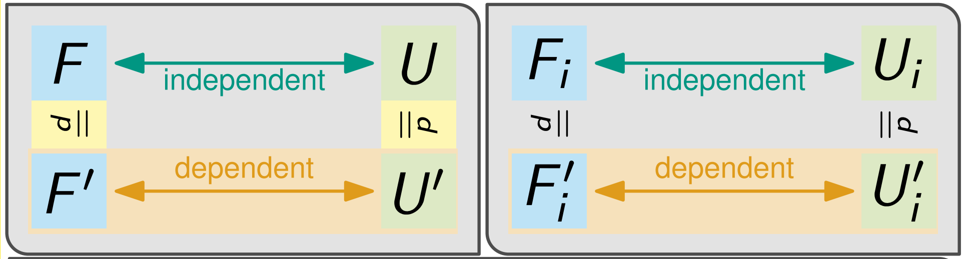
$F = \sum_{i=1}^{n} F_i$

$U = \sum_{i=1}^{n} U_i$

$F' = \sum_{i=1}^{n} F'_i$

$U' = \sum_{i=1}^{n} U'_i$

Coupling: Random variables X_1, X_2. Define random variables X'_1, X'_2 in a shared probability space such that $X_1 \overset{d}{=} X'_1$ and $X_2 \overset{d}{=} X'_2$.



Observation: Independent rand. var. X_i, Y_i for $i \in [n]$ with couplings (X'_i, Y'_i) for $i \in [n]$.
Then, for any function f: $(f(X'_1, \dots, X'_n), f(Y'_1, \dots, Y'_n))$ is a coupling of $f(X_1, \dots, X_n)$ and $f(Y_1, \dots, Y_n)$.

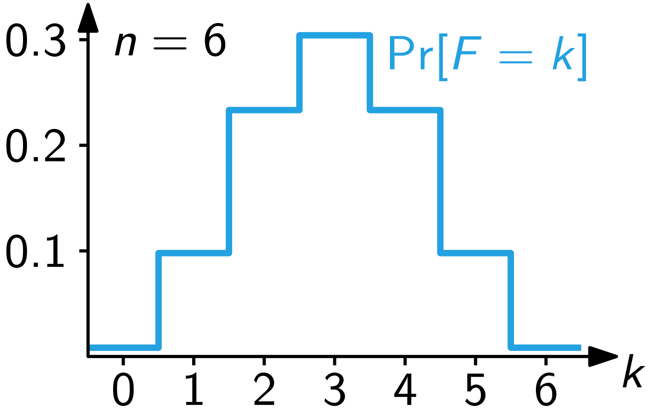
The Binomial-Poisson-Approximation or "How I Lied To You"

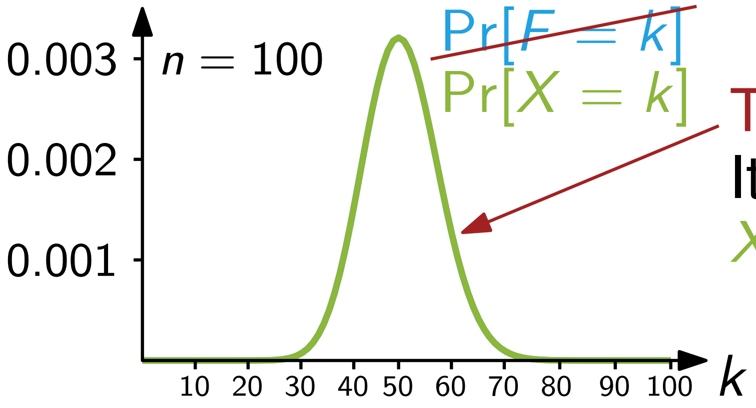
Setup

- Fair $\{0,1\}$-coin X with $\Pr[X=1]=p=\frac{1}{2}$ — This is a Bernoulli rand. var. $X \sim \mathrm{Ber}(p)$
- Sum of n ind. coins $F = \sum_{i=1}^{n} X_i$, $X_i \sim \mathrm{Ber}(p)$ — This is a Binomial rand. var. $F \sim \mathrm{Bin}(n,p)$

$$\Pr[F=k] = \binom{n}{k} p^k (1-p)^{n-k}$$

- ... which we have seen today already





This is not a binomial distribution!
It's a Poisson distribution with $\lambda = 50$
$X \sim \mathrm{Pois}(\lambda)$: $\Pr[X=k] = \lambda^k e^{-\lambda}/k!$

- Why lie? It was easier to plot that way and I thought you wouldn't notice...
- How dare I? As n increases, the two distributions are very close...

What does that mean?

Total Variation Distance

- A measure of distance between the distributions of random variables

(Disclaimer: In the following we use a very simplified notation that abstracts away a lot of details!)

Definition: Let X, Y be random variables taking values in a set S. The **total variation distance** of X and Y is $d_{TV}(X,Y) = \frac{1}{2}\sum_{x\in S} |\Pr[X=x] - \Pr[Y=x]|$.

- Intuition: Sum over the differences in the probabilities
- Maybe a bit tedious to work with, simple bound:

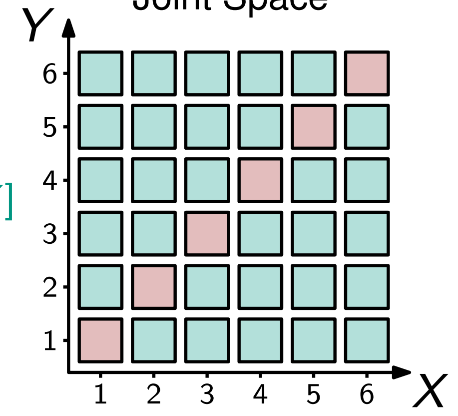
Fréchet: $\Pr[A] - \Pr[B] \leq \Pr[A \wedge \bar{B}]$

$S_X = \{x \in S \mid \Pr[X=x] \geq Pr[Y=x]\}$ $\quad S_Y = S \setminus S_X$

$$
\begin{aligned}
\cancel{2}d_{TV}(X,Y) &= \sum_{x\in S} |\Pr[X=x] - \Pr[Y=x]| \\
&= \sum_{x\in S_X} \cancel{|}\Pr[X=x] - \Pr[Y=x]\cancel{|} + \sum_{x\in S_Y} \cancel{|}\Pr[X=x] - \Pr[Y=x]\cancel{|} \\
&= \sum_{x\in S_X} \Pr[X=x] - \Pr[Y=x] + \sum_{x\in S_Y} \Pr[Y=x] - \Pr[X=x] \\
&\leq \sum_{x\in S_X} \Pr[X=x \wedge Y\neq x] + \sum_{x\in S_Y} \Pr[Y=x \wedge X\neq x] \\
&\leq \sum_{x\in S} \Pr[X=x \wedge Y\neq x] + \sum_{x\in S} \Pr[Y=x \wedge X\neq x] \\
&= \Pr[X\neq Y] + \Pr[Y\neq X] = \cancel{2}\Pr[X\neq Y]
\end{aligned}
$$

Joint Space

$\Pr[X\neq Y]$



Lemma: $d_{TV}(X,Y) \leq \Pr[X \neq Y]$.

Total Variation Distance

Karlsruhe Institute of Technology

- A measure of distance between the distributions of random variables

(Disclaimer: In the following we use a very simplified notation that abstracts away a lot of details!)

Definition: Let X, Y be random variables taking values in a set S. The **total variation distance** of X and Y is $d_{TV}(X,Y) = \frac{1}{2}\sum_{x \in S} |\Pr[X = x] - \Pr[Y = x]|$.

- Intuition: Sum over the differences in the probabilities
- Maybe a bit tedious to work with, simple bound:

Fréchet: $\Pr[A] - \Pr[B] \leq \Pr[A \wedge \bar{B}]$

Lemma: $d_{TV}(X,Y) \leq \Pr[X \neq Y]$.

- Note that d_{TV} is defined via the distributions of X and Y
- For any coupling (X', Y') of X, Y we have $X' \overset{d}{=} X$ and $Y' \overset{d}{=} Y$. Thus, $d_{TV}(X,Y) = d_{TV}(X',Y')$

Lemma (coupling inequality): Let X, Y be random variables. Then for any coupling (X', Y') of X and Y it holds that $d_{TV}(X,Y) \leq \Pr[X' \neq Y']$.

Lemma (triangle inequality): For rand. var. X, Y, Z: $d_{TV}(X,Z) \leq d_{TV}(X,Y) + d_{TV}(Y,Z)$.

The Binomial-Poisson-Approximation

- Ind. $X_i \sim \mathrm{Ber}(p)$ for $i \in [n]$ $\longrightarrow$ $X = \sum_{i=1}^n X_i$ $\longrightarrow$ $X \sim \mathrm{Bin}(n,p)$
- Ind. $Y_i \sim \mathrm{Pois}(\lambda)$ for $i \in [n]$, $\lambda = -\log(1-p)$ $\longrightarrow$ $Y = \sum_{i=1}^n Y_i$ $\longrightarrow$ $Y \sim \mathrm{Pois}(n\lambda)$

Lemma: $X \sim \mathrm{Bin}(n,p)$, $Y \sim \mathrm{Pois}(-n\log(1-p))$: $d_{TV}(X,Y) \leq \frac{n}{2}\log(1-p)^2$.

$\Pr[Y_i = k] = e^{-\lambda}\lambda^k/k!$

$$\begin{array}{ccc} X_i & \xleftrightarrow{\text{independent}} & Y_i \\ \overset{d}{=}\ \checkmark & & \overset{d}{=}\ \checkmark \\ X_i' & \xleftrightarrow{\text{dependent}} & Y_i' \end{array}$$

Proof

- For each i we couple Y_i and X_i: $Y_i' = Y_i$, $X_i' = \min\{Y_i', 1\}$
- To show that this is a coupling, we need $X_i \overset{d}{=} X_i'$

$$\Pr[X_i' = 0] = \Pr[Y_i = 0] = e^{-\lambda} = e^{\log(1-p)} = 1 - p = \Pr[X_i = 0] \checkmark$$

$$\Pr[X_i' = 1] = \Pr[Y_i > 0] = 1 - \Pr[Y_i = 0] = 1 - \Pr[X_i = 0] = \Pr[X_i = 1] \checkmark$$

- $X' = \sum_{i=1}^n X_i'$, $Y' = \sum_{i=1}^n Y_i'$ $\Rightarrow$ (X', Y') coupling of (X, Y)

$$d_{TV}(X,Y) \leq \Pr[X' \neq Y'] \underset{\text{union bound}}{\leq} \sum_{i=1}^n \Pr[X_i' \neq Y_i'] = \sum_{i=1}^n \Pr[Y_i' \geq 2]$$

$$= \sum_{i=1}^n e^{-\lambda} \sum_{j \geq 2} \frac{\lambda^j}{j!} \leq \sum_{i=1}^n \frac{\lambda^2}{2} \cancel{e^{-\lambda}} \underbrace{\cancel{\sum_{j \geq 0} \frac{\lambda^j}{j!}}}_{= e^{\lambda}} = \sum_{i=1}^n \frac{\lambda^2}{2} \checkmark$$

Function of Couplings
Couplings (X_i', Y_i') of (X_i, Y_i):
$(f(X_i'), f(Y_i'))$ coupling of $(f(X_i), f(Y_i))$.

Coupling Inequality
For any coupling (X', Y') of X, Y:
$d_{TV}(X,Y) \leq \Pr[X' \neq Y']$.

Recap: Theory-Practice Gap

Karlsruhe Institute of Technology

Theory

- Many computational problems are assumed to be hard
- Looks like there are no algorithms that can solve these problems fast

$\underbrace{\text{SAT} \quad \text{Vertex Cover} \quad \text{Independent Set}}_{\text{NP-hard}}$

Practice

- Many computational problems can be solved extremely fast
  - “Modern SAT solvers can often handle problems with millions of clauses and hundreds of thousands of variables”

    “Propagation = Lazy Clause Generation”, Ohrimenko, Stuckey & Codish, CP, 2017
  - For many real-world graphs optimal vertex covers (containing up to millions of nodes) can be found in seconds

    “Branch-and-reduce exponential/FPT algorithms in practice: A case study of vertex cover”, Akiba & Iwata, TCS, 2016

Average-Case Analysis

- Acknowledge difference between theoretical worst-case instances and practical ones
- Represent real world using mathematical models and analyze those theoretically

Random Graph Models

A **graph model** describes a mechanism that can be used to generate a graph.

- Given a set of vertices, how are edges in the graph formed?
- The model consists of *rules* defining which vertices are adjacent
- In *random* graph models these rules involve randomness

Desirable Properties

- *Simplicity*: We cannot analyze a model that is too complicated
- *Realism*: We do not want to analyze a model that cannot be used to make predictions about the real world
- *Fast Generation*: We want to be able to generate many, large benchmark instances to . . .
  - analyze structural and algorithmic properties empirically
  - generate hypotheses about asymptotic behavior

Let's start with a simple model!

Erdős–Rényi Random Graphs

Karlsruhe Institute of Technology

History

- Initially introduced by Edgar Gilbert in 1959

"Random Graphs", Gilbert, Ann. Math. Statist., 1959

- A related version introduced by Paul Erdős and Alfréd Rényi in 1959

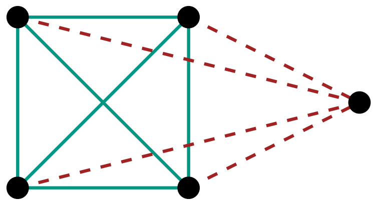
"On Random Graphs I", Erdős & Rényi, Publ. Math. Debr., 1959

Definitions

Gilbert's model, though often meant when talking about Erdős–Rényi graphs

$\boldsymbol{G(n,p)}$	$\boldsymbol{G(n,m)}$
■ Start with n nodes ■ Independently connect any two with fixed probability p	■ Start with n nodes ■ From the $\binom{n}{2}$ possible edges select m uniformly at random

- For $\tilde{p} = m/\binom{n}{2}$ the *expected* number of edges in $G(n, \tilde{p})$ matches m
- In $G(n,p)$ edges are independent, in $G(n,m)$ they are not
  - If a $G(5,6)$ contains a 4-clique, there can be no edge incident to the 5th node



Existence of red edges depends on existence of green ones.

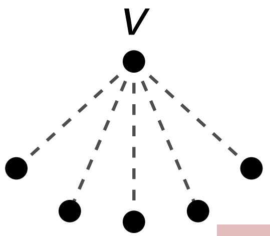
- Since many real-world networks are $\overbrace{\textit{sparse}}^{\text{number of edges linear in number of nodes}}$, we focus on $p = \frac{c}{n}$ for $c \in \Theta(1)$

ER – Degree of a Vertex

Karlsruhe Institute of Technology

Vertex Degree

- Number of neighbors, number of incident edges
- each of $n-1$ potential edges exists with prob. p
- $\deg(v) \sim \mathrm{Bin}(n-1, p)$ → $\Pr[\deg(v) = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}$
  → $\mathbb{E}[\deg(v)] = (n-1)p$ & $\mathrm{Var}[\deg(v)] = (n-1)p(1-p)$ *Inconvenient...*



$G(n,p)$
Independently connect any two nodes with fixed probability p.

$p = \frac{c}{n}$, $c \in \Theta(1)$

Approximation

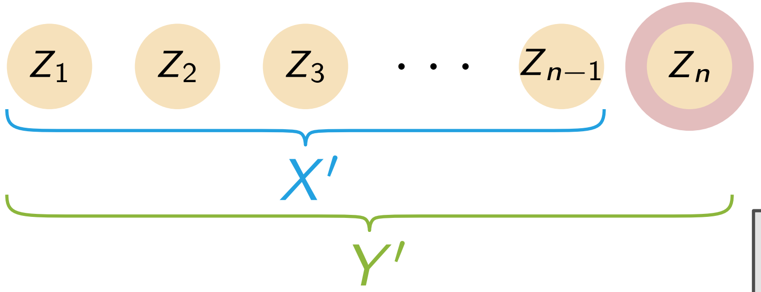
Lemma: Let $p = \frac{c}{n}$ for $c \in \Theta(1)$, let $X \sim \mathrm{Bin}(n-1, p)$ and let $Y \sim \mathrm{Bin}(n, p)$. Then, $d_{TV}(X, Y) = o(1)$.

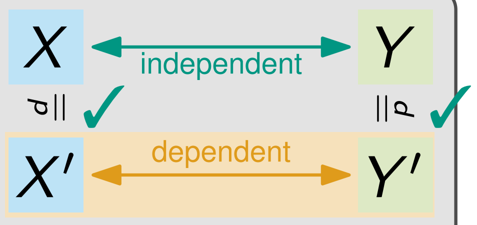
Proof

- Independent $Z_i \sim \mathrm{Ber}(p)$ for $i \in [n]$
- $X' = \sum_{i=1}^{n-1} Z_i$, $Y' = X' + Z_n$

$d_{TV}(X, Y) \le \Pr[X' \neq Y']$

$= \Pr[Z_n = 1] = \frac{c}{n} = o(1)$ ✓



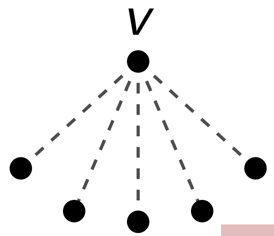


Coupling Inequality
For any coupling (X', Y') of X, Y:
$d_{TV}(X, Y) \le \Pr[X' \neq Y']$.

ER – Degree of a Vertex

Vertex Degree

- Number of neighbors, number of incident edges
- each of $n-1$ potential edges exists with prob. p
- $\deg(v) \sim \mathrm{Bin}(n-1, p)$ → $\Pr[\deg(v) = k] = \binom{n-1}{k} p^k (1-p)^{n-1-k}$
  → $\mathbb{E}[\deg(v)] = (n-1)p$ & $\mathrm{Var}[\deg(v)] = (n-1)p(1-p)$ *Inconvenient...*



$G(n,p)$
Independently connect any two nodes with fixed probability p.

$p = \frac{c}{n}$, $c \in \Theta(1)$

Approximation

Lemma: Let $p = \frac{c}{n}$ for $c \in \Theta(1)$, let $X \sim \mathrm{Bin}(n-1, p)$ and let $Y \sim \mathrm{Bin}(n, p)$. Then, $d_{TV}(X, Y) = o(1)$. And for $Z \sim \mathrm{Pois}(c + O(\frac{1}{n}))$: $d_{TV}(X, Z) = o(1)$.

$$\begin{aligned} d_{TV}(X,Z) &\le d_{TV}(X,Y) + d_{TV}(Y,Z) \\ &= o(1) + \underbrace{\tfrac{n}{2}\log(1-p)^2}_{} = o(1) \end{aligned}$$

$$\begin{aligned} \tfrac{n}{2}(-p - O(p^2)))^2 &= \tfrac{n}{2}(p^2 + O(p^3)) \\ &= \tfrac{n}{2}((\tfrac{c}{n})^2 + O((\tfrac{c}{n})^3)) \\ &= \tfrac{c^2}{2n} + O(\tfrac{c^3}{n^2}) = o(1) \end{aligned}$$

- $\mathbb{E}[Z] = \mathrm{Var}[Z] \approx c$, much simpler than the above!

Binomial-Poisson-Approximation
$Y \sim \mathrm{Bin}(n, p)$, $Z \sim \mathrm{Pois}(-n\log(1-p))$:
$d_{TV}(Y, Z) \le \frac{n}{2}\log(1-p)^2$.

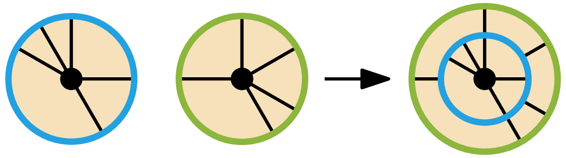
Triangle Inequality
$d_{TV}(X, Z) \le d_{TV}(X, Y) + d_{TV}(Y, Z)$.

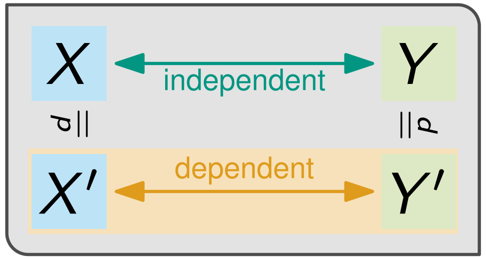
Taylor $p \to 0$: $\log(1-p) = -p - O(p^2)$

Conclusion

Coupling

- Define relation between rand. var. to make statements about one using the other
- A coupling of (X, Y) is a pair (X', Y') of random variables in a shared probability space such that $X \overset{d}{=} X'$ and $Y \overset{d}{=} Y'$
- Often X' and Y' dependent
- Examples: Wheel of fortune & Unfair dice
- Coupling inequality to bound *total variation distance*





Random Graph Models

- Mathematical models represent real-world networks and allow for theoretical analysis
- Desirable properties: simple, realistic, fast to generate

Erdős-Rényi Random Graphs

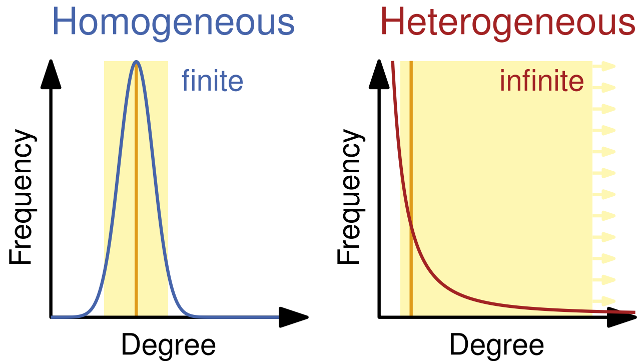
- $G(n, p)$: Start with n nodes, connect any two with fixed probability p, independently
- In sparse $G(n, p)$ the degree of a vertex is approximately Poisson-distributed

Outlook: Degree Distribution vs. Degree Distribution

Distributions

- Probability distribution of the degree of a *given* vertex in a $G(n, \frac{c}{n})$ approaches $\mathrm{Pois}(c)$
- Empirical distribution of the degrees of *all* vertices in a graph $G = (V, E)$

$N_d = \sum_{v \in V} \mathbb{1}_{\{\deg(v) = d\}}$ (normalized: $\frac{1}{n} N_d$, for $n = |V|$)



Characterizing a Distribution

- *Mean*: What degree would we expect for a vertex?
- *Variance*: (very rough intuition) How far would we expect the degree of a vertex to deviate from the mean?

Empirical Distribution of $G(n, \frac{c}{n})$ → homogeneous